TOWARD AUTOMATIC RECOGNITION OF JAPANESE BROADCAST NEWS

Tatsuo Matsuoka†, Yuichi Taguchi††, Katsutoshi Ohtsuki†, Sadaoki Furui†, and Katsuhiko Shirai††

† NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180, Japan
†† Department of Information and Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169, Japan
e-mail: matsuoka@splab.hil.ntt.co.jp

ABSTRACT

In this paper we report on automatic recognition of Japanese broadcast-news speech. We have been working on large-vocabulary continuous speech recognition (LVCSR) for Japanese newspaper speech transcription and achieved reasonably good performance. We have recently applied our LVCSR system to transcribing Japanese broadcast-news speech. We extended the vocabulary to 20k words and trained the language models using newspaper texts and broadcast-news manuscripts. These two language models were applied to our evaluation speech sets. The language model trained using broadcast-news manuscripts achieved better results for broadcast-news speech than the language model trained using newspaper texts which achieved better results for newspaper speech. In preliminary experiments on Japanese broadcast-news transcription, we achieved a word accuracy of 79.3% for anchor-speakers' speech by using a language model trained using broadcast-news manuscripts and newspaper texts.

1. INTRODUCTION

The DARPA Hub-4 test that began in 1995 is evaluating the use of large-vocabulary continuous speech recognition (LVCSR) to transcribe audio recordings of broadcast news. Several preliminary Hub-4 evaluation results have been reported [1-5]. Coincidentally, in 1996, the Japanese government announced that it will issue a regulation in several years requiring TV news programs to be closed captioned. Transcribing broadcast news is a challenging task, and thus a good test of applying LVCSR technology to real-world systems. We are therefore investigating the automatic recognition of Japanese broadcast-news speech. This paper describes some of our preliminary results.

We have been working on LVCSR for read newspaper speech. So far, a word accuracy of about 90% has been achieved for a 7k-word-vocabulary [6-8]. Figure 1 shows the progress of our LVCSR performance for newspaper speech recognition. We found that bigram and trigram language models are very effective for Japanese LVCSR. Our trigram language model reduced the word error rate from 18.1% to 10.1%. This improvement is much larger than those for other languages.

We have applied our LVCSR system to transcribing broadcast-news speech. We extended the vocabulary to 20k words and trained the language models using newspaper texts and broadcast-news manuscripts. We conducted phoneme-recognition experiments to examine if broadcast-news speech is acoustically more difficult than read newspaper speech. Then we experimentally compared two language models: one trained using broadcast-news manuscripts and one trained using newspaper texts.

2. BROADCAST NEWS DATA

Raw audio recordings of broadcast news include frequent speaker changes, background music, and telephone speech, such as field reports. We segmented these parts manually and used only the clean-speech parts, i.e., those parts not containing background music, noise, or telephone speech for the experiments reported here. Even using only clean speech is still challenging because news speech is usually much more fluent than read speech. Furthermore, we found that the sentences are much longer for broadcast news than for newspapers. As shown in Fig. 2, the average number of words

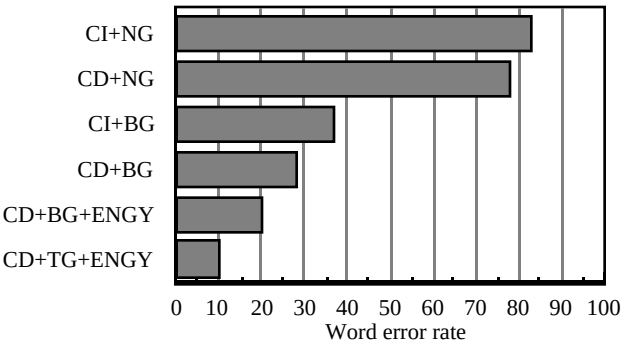

Fig. 1 LV CSR experimental results for newspaper speech.
CI: context-independent acoustic models were used, CD: context-dependent acoustic models were used, NG: no grammar models were used, BG: bigram language models were used, TG: trigram language models were used, ENGY: energy parameters were added to the feature parameters.

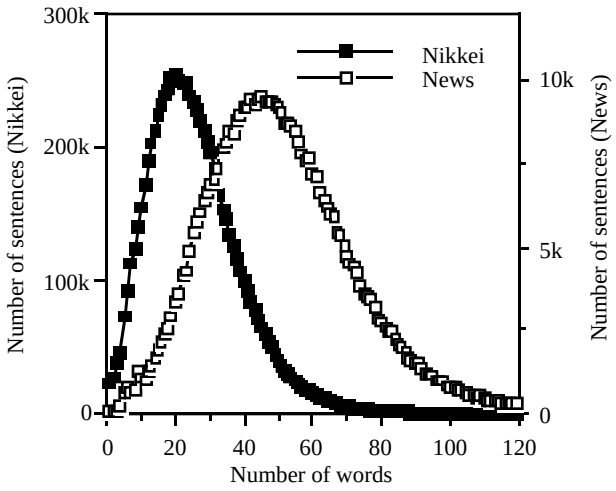

Fig. 2 Histogram of the number of words per sentence

per sentence in broadcast-news manuscripts is almost double that in newspaper texts.

To apply n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer because Japanese sentences are written without spacing between words. Some of the irrelevant symbol-marks, such as bullets, were filtered out. The details of this text filtering process are described in [6, 7]. A word-frequency list was derived from the filtered sentences, and the 20k most frequently used words were selected as the vocabulary words. This 20k-vocabulary covers about 98% of the words in the broadcast-news manuscripts. Table 1 lists the training-text size and the coverage for broadcast-news, the Nikkei newspaper and the Wall Street Journal. The training-text size is noticeably smaller for the broadcast news than for the newspapers.

3. ACOUSTIC MODELING

The acoustic models we used were all shared-state context-dependent phoneme HMMs designed using tree-based clustering [9]. The total number of states was 2106, and the number of Gaussian mixture components per state was 4. They were trained using phonetically-balanced sentences and read dialogue speech spoken by 53 speakers. The total number of utterances was 13,270.

To investigate the acoustical difference between broadcast-news speech and read newspaper speech we conducted phoneme-recognition experiments. Table 2 shows the phoneme recognition results. The percent correct and accuracy were calculated as follows.

$$\%Correct = \frac{N - sub. - del.}{N} \cdot 100$$

Table 1 Comparison of lexica and LM training

	News	Nikkei	WSJ
Training text size (words)	24M	180M	237M
Number of distinct words	114k	623k	476k
5k coverage	91.5%	88.0%	90.6%
7k coverage	-	90.3%	-
20k coverage	98.0%	96.2%	97.5%
30k coverage	-	97.5%	-
65k coverage	99.7%	99.0%	99.6%

Table 2 Phoneme recognition for news speech and read newspaper speech

	News	Nikkei
% Correct	82.0%	80.7%
Accuracy	61.5%	64.7%

Table 3 Number and average occurrence of distinct n-grams

	n-gram	Distinct no.	Av. occurrence
Broadcast news	unigram	20k	1160
	bigram	0.9M	24
Nikkei newspaper	unigram	20k	8747
	bigram	3.6M	44

Table 4 Evaluation speech

	Anchor	Others	Nikkei
No. of speakers	5	6	10
No. of utterances	100	125	100
No. of words	4184	2285	2168
OOV rate	0.9%	3.7%	3.5%

$$Accuracy = \frac{N - sub. - del. - ins.}{N} \cdot 100$$

The accuracy was almost the same for the broadcast-news speech and the read newspaper speech. Therefore, it can be said that acoustic models trained using read speech are applicable to news speech LVCSR.

4. LANGUAGE MODELING

N-gram language models have shown significant effectiveness in Japanese LVCSR for read newspaper speech [6-8]. We can expect the same effectiveness for news speech transcription. To train n-gram language models we need a large amount of text data. As Table 1 shows, it is usually easier to collect a large amount of data for newspaper texts than for broadcast-news manuscripts. Therefore, it would be helpful if a newspaper language model also worked well for broadcast news.

To determine if a newspaper language model can be used for broadcast news, we trained two language models, one using

Table 5 LVCSR results

Task	Language model	Test-set perplexity	Word accuracy
News (Anchor)	News LM	105	76.3
	Nikkei LM	190	68.5
News (Others)	News LM	255	61.8
	Nikkei LM	281	59.8
Nikkei (30k)	News LM	253	69.0
	Nikkei LM	100	77.2

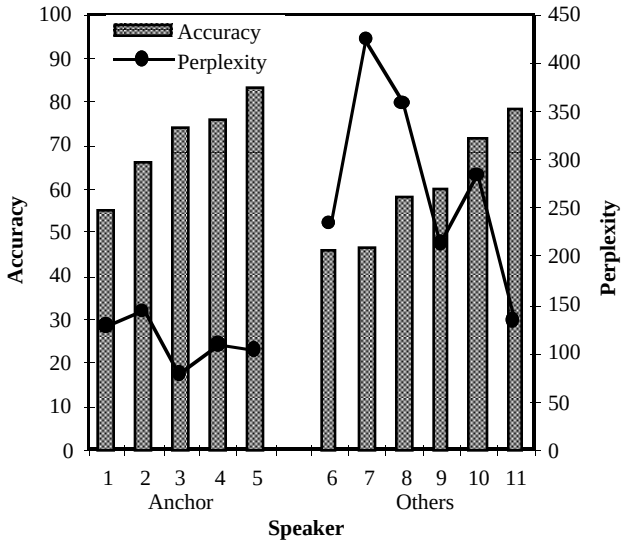

Fig. 3 LVCSR results

broadcast-news manuscripts and one using newspaper texts. Table 3 shows the number of distinct unigrams and bigrams and the average occurrence of n-gram models in the training texts. The number of distinct bigrams for broadcast news was much smaller than that for the Nikkei newspaper due to the small training-text size.

The language models used in the LVCSR experiments described in the next section were basically bigram models unless otherwise described. The bigram models were smoothed using Katz's smoothing method [10].

5. LVCSR EXPERIMENTS

The evaluation speech sets are summarized in Table 4. The news-speech data set was divided into two parts: one for anchor speakers and one for other speakers. For comparison, we also used a read newspaper speech set that had a 30k-vocabulary.

The LVCSR results are shown along with the test-set perplexities in Table 5. The two language models (LMs) were applied to each evaluation speech set. The News LM achieved better results for news speech (Anchor and Others) than the Nikkei LM, which achieved better results for read newspaper

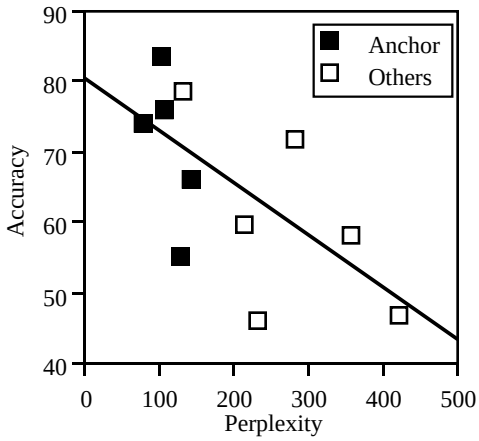

Fig. 4 Perplexity vs. word accuracy

speech (Nikkei). To investigate why the News LM showed poor performance for Others, we plotted word accuracy for each speaker against test-set perplexity (Figs. 3 and 4). We found that word accuracy depends on test-set perplexity, not on the type of speaker (See Fig. 4).

Since we had few broadcast-news manuscripts, we trained a trigram model not using broadcast-news manuscripts but using newspaper texts and applied it to broadcast-news speech transcription. The word accuracy improved from 76.3% to 79.3% for Anchor speakers' speech. This result suggests that word trigram models will be reasonably effective for broadcast-news speech recognition if there is a large amount of text data from the same task domain.

6. CONCLUSION

In preliminary experiments on Japanese broadcast-news transcription, we have achieved a word accuracy of 76.3% for anchor-speakers' speech by using a bigram language model trained using broadcast-news manuscripts. A trigram language model trained using newspaper texts improved the word accuracy to 79.3%. Phoneme-recognition experiments showed that acoustic models trained using read speech are applicable to broadcast-news speech. A newspaper language model was not as effective as a broadcast-news language model for broadcast-news transcription. A language model interpolation or adaptation method is definitely needed.

We are currently working on a language model adaptation that will enable us to use a large number of newspaper texts for training a language model for broadcast-news transcription. We are also incorporating trigram language models to our broadcast-news transcription system.

References

1. F. Kubala, T. Anastasakos, H. Jin, J. Makhoul, L. Nguyen, R. Schwartz, and N. Yuan, "Toward automatic

recognition of broadcast news," Proc. DARPA Speech Recognition Workshop, pp. 55-60, February 1996

2. U. Jain, M. A. Siegler, S. J. Doh, E. Gouvea, J. Huerta, P. J. Moreno, B. Raj, and R. M Stern, "Recognition of continuous broadcast news with multiple unknown speakers and environments," Proc. DARPA Speech Recognition Workshop, pp. 61-66, February 1996

3. S. Wegmann, et al., "Marketplace recognition using Dragon's continuous speech recognition system," Proc. DARPA Speech Recognition Workshop, pp. 67-71, February 1996

4. P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, L. Polymenakos, H. Printz, and M. Franz, "Transcription of radio broadcast news with the IBM large vocabulary speech recognition system," Proc. DARPA Speech Recognition Workshop, pp. 72-76, February 1996

5. F. Kubala, T. Anastasakos, H. Jin, L. Nguyen, and R. Schwartz, "Transcribing radio news," ICSLP-96, pp. 598-601, October 1996

6. T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Large-vocabulary continuous speech recognition using a Japanese business newspaper (Nikkei)," DARPA Speech Recognition Workshop, pp. 137-142, February 1996

7. T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Japanese large-vocabulary continuous speech recognition using a business-newspaper corpus," ICSLP-96, pp. 22-25, October 1996

8. T. Matsuoka, K. Ohtsuki, T. Mori, K. Yoshida, S. Furui, and K. Shirai, "Japanese large-vocabulary continuous speech recognition using a business-newspaper corpus," ICASSP-97, to appear, April 1997

9. S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. ARPA Human Language Technology Workshop, pp. 307-312, March 1994

10. S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," Trans. ASSP-35, pp. 400-401, March 1987